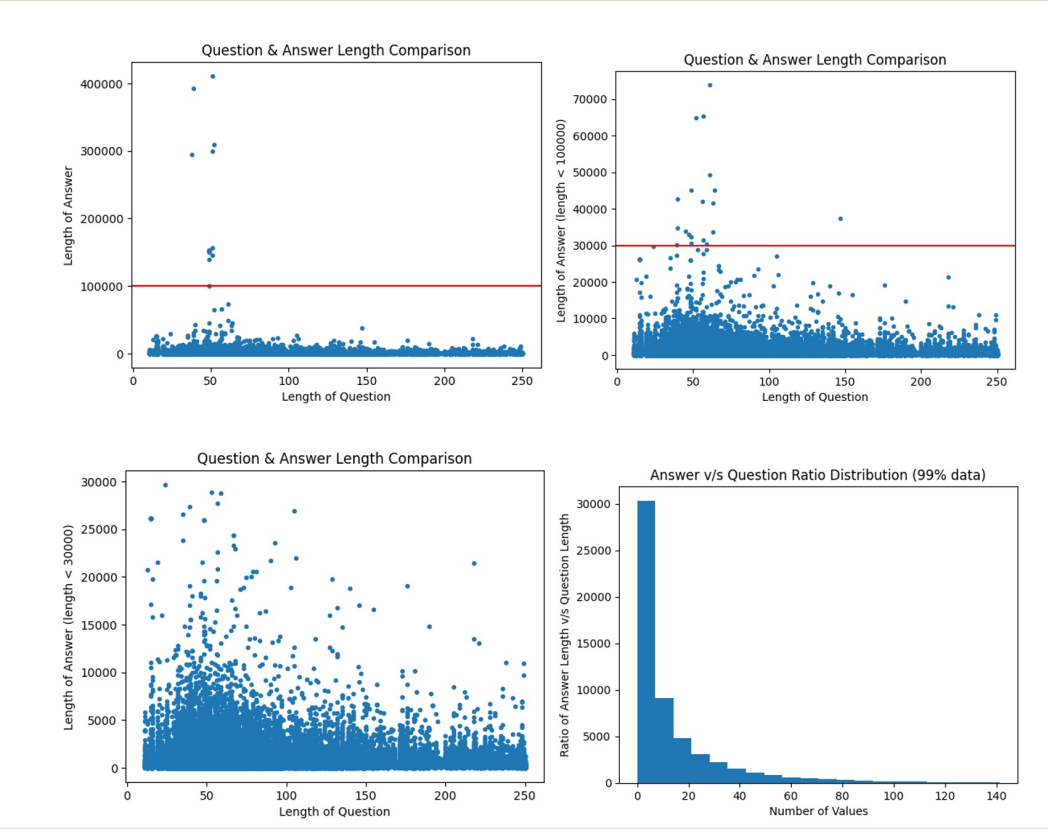
HACK TO HIRE

Srishti Sahni

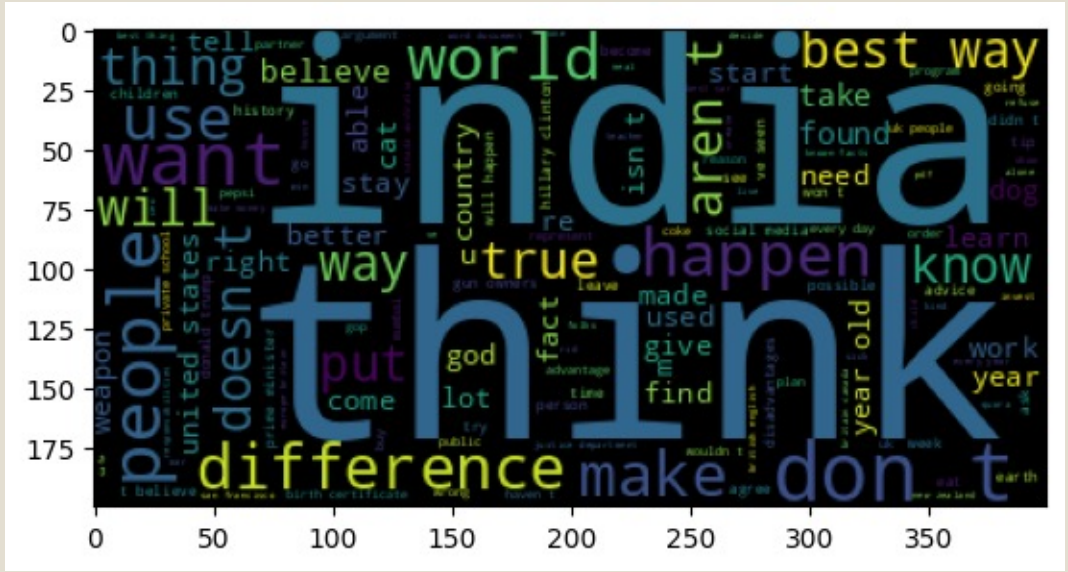
Libraries Used

Library	Use
pandas	Data manipulation and analysis
transformers	State-of-the-art NLP models
torch	Deep learning
nltk.translate.bleu_score	BLEU score calculation
rouge_score	ROUGE score calculation
sklearn.metrics	Evaluating machine learning models
numpy	Numerical computations
matplotlib.pyplot	Creating visualizations
wordcloud	Creating word clouds
nltk	Natural language processing tools
nltk.corpus.stopwords	Excluding common words from text analysis
nltk.stem.PorterStemmer	Stemming words to their root form
string	Common string operations

Exploratory Data Analysis



Length Analysis



Word Frequency Using Word Clouds

Model Architecture

BERT	GPT-2	T5
• Pre-trained on a large corpus and fine-tuned on the Quora dataset. • Used for extracting contextual embeddings of questions and answers. • Architecture involves feeding the [CLS] token representation to a classifier for predicting the start and end positions of the answer.	• Pre-trained on a large text corpus and fine-tuned on the Quora dataset. • Utilizes a decoder-only architecture that generates answers based on the given question. • Fine-tuning involves adjusting the model to minimize the loss on the Quora question-answer pairs.	• Converts all NLP tasks into a text-to-text format. • Fine-tuned by framing the question-answering task as text generation. • Input format: "question: <question> • Output format: "<answer>"

Training Process

- Hyperparameter tuning was performed for each model to find the optimal learning rate, batch size, and number of epochs.
- The Adam optimizer was used for training with a learning rate scheduler to adjust the learning rate during training

Evaluation Metrics

- **F1 Score:** Measures the overlap between the predicted and ground truth answers, considering both precision and recall.
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** Measures the overlap of n-grams between the predicted and ground truth answers, focusing on recall.
- **BLEU (Bilingual Evaluation Understudy):** Measures the precision of n-grams in the predicted answer compared to the ground truth, focusing on precision.

Results

Model Name	BLEU	ROUGE-L	F1-Score
BERT	0.0002	0.02517	0.02546
GPT-2	0.002	0.088	0.093
T5	-	-	-

Conclusion

- **GPT-2 demonstrated superior performance over BERT.**
  - Expected outcome as BERT was trained without any context.
  - BERT needs context to develop Question/Answer models.
  - GPT-2 is more suited to text-generation problems.
- **T5 offers flexibility and can be trained for various problem-sets.**
  - Not compared in this context.
  - Expected to perform similarly or better than GPT-2.

Future Work (Novel Suggestions)

- Experimenting with ensemble methods to combine the strengths of these models and further improving the pre-processing and fine-tuning processes.
- Using information retrieval techniques to extract context for the questions being asked.
- The goal was to create human like answers, to achieve that, training a custom RNN with a larger data-set and custom embeddings can help achieving that task.